Explainable Span-level Hallucination Detection Dataset Construction via Iterative Self-refinement

Anonymous ACL submission

Abstract

Effective span-level hallucination detection is crucial for the trustworthy deployment of Large Language Models (LLMs), relying on high-quality datasets as its foundational prerequisite. To alleviate the high cost of span-level manual annotation, prior work has turned to automatic dataset construction. Existing automatically constructed span-level datasets are typically created by synthesizing hallucinated claims via span perturbations with label annotations. However, such synthetic hallucinated claims often contain lots of surface-level flaws such as disfluency, incoherence, or inconsistency, which differ from the fluent, coherent, consistent yet factually incorrect hallucinations produced by real-world LLMs; meanwhile, label-only annotations lack explainability and practical utility. To tackle the above problems, in this paper, we propose a novel self-refinement framework designed to generate more realistic hallucinated claims and explainable annotations by leveraging structured hallucination-inducing edit suggestions and a generate–critique–refine loop mechanism, resulting in a high-quality span-level dataset with explainable annotations, named ExpHul. Moreover, we train a compact hallucination detection model on ExpHul. Extensive experiments show that the detection model outperforms state-of-the-art baselines while providing high-quality explanations and revisions. Furthermore, the 8B-parameter detection model achieves competitive performance with the best closed-source LLMs like Gemini-3.0-Pro and GPT-5.2 at a lower inference cost.

1 Introduction

Large language models (LLMs) have achieved impressive performance across a wide range of tasks, yet they remain prone to hallucinations, producing content that appears plausible but is factually incorrect or unsupported (Bao et al., 2025; Bazarova et al., 2025; Ji et al., 2024). Such hallucinations

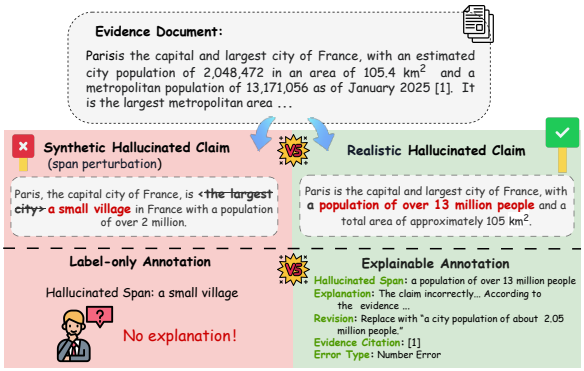

Figure 1: Comparison of hallucination generation and annotation. The top row contrasts synthetic hallucinated claims via span-level perturbation with realistic hallucinated claims. The bottom row compares label-only annotation with our explainable annotation, which augments hallucinated spans with explanations, evidence, revisions, and error types.

undermine the reliability and safety of LLMs, particularly in high-stakes domains such as finance, law, and medicine (Bao et al., 2025; Ji et al., 2023). Thus, accurately detecting hallucinations is critical for trustworthy LLM deployment, highlighting the need for more fine-grained hallucination detection. Compared with coarse-grained response-level hallucination detection, span-level hallucination detection is more promising, as it can pinpoint the exact hallucinated spans, thereby enabling more precise error diagnosis and more targeted mitigation strategies.

Existing span-level hallucination detection datasets can be divided into two categories: manually annotated and automatically constructed. Manually annotated datasets (Niu et al., 2024) require annotators to compare LLM-generated claims against lengthy evidence documents and identify fine-grained hallucinated spans, which is labor-intensive, costly, and difficult to scale. In contrast, automatically constructed datasets offer better scalability and cost efficiency. Existing automatically

constructed datasets (Mishra et al., 2024) typically synthesize hallucinations via span perturbations, which generate hallucinated claims by replacing or inserting spans into factually correct claims, and provide the corresponding span-level hallucination categorical labels.

However, due to synthesizing hallucinations via span perturbations, these automatically constructed datasets tend to contain hallucinated claims with surface-level errors such as disfluency, incoherence, or inconsistency, which differ from the fluent, coherent, and consistent yet factually incorrect hallucinations generated by real LLMs. Thus, the hallucinated spans can often be identified using shallow cues without requiring deep semantic understanding and factual reasoning over the evidence documents and the hallucinated claims. Models trained on such datasets may overfit to these low-quality synthetic patterns and fail to generalize to realistic hallucinations in real-world LLM outputs. Furthermore, existing automatically constructed datasets only provide span-level hallucination categorical labels without explanations, preventing models from learning why a span is hallucinated and limiting the explainability and trustworthiness of downstream detection systems. As shown in Figure 1, the synthetic claim contains an inconsistency ("small village" vs. "capital city" and "over 2 million"), making the hallucination trivially detectable without evidence comparison. The label-only annotation only contains a hallucinated span, without an explanation clarifying why the span is hallucinated.

To tackle the above problems, in this paper, we propose a novel self-refinement framework designed to generate more realistic hallucinated claims and explainable annotations by leveraging structured hallucination-inducing edit suggestions and a generate–critique–refine loop mechanism. Specifically, the structured hallucination-inducing edit suggestions explicitly encode hallucination types, original correct spans, and corresponding hallucinated revisions, which can guide both hallucinated claim synthesis and explainable annotation. Moreover, the generate–critique–refine loop iteratively evaluates and refines the edit suggestions based on the fluency, coherence, and consistency of the generated hallucinated claims, yielding natural yet factually incorrect hallucinations that better reflect real LLM outputs. Leveraging this framework, we construct ExpHul, an explainable span-level hallucination detection dataset with comprehensive annotations, including hallucinated spans, explanations, revision suggestions, supporting evidences, and fine-grained labels.

Furthermore, we train an 8B-parameter hallucination detection model on ExpHul. Extensive experiments demonstrate that the detection model outperforms state-of-the-art baselines while providing high-quality explanations and revisions. Moreover, the 8B-parameter hallucination detection model achieves competitive performance with the best closed-source LLMs like Gemini-3.0-Pro and GPT-5.2 at a lower inference cost.

2 Related Work

Existing hallucination detection methods can be mainly categorized into response-level and span-level approaches. Response-level methods treat hallucination detection as a binary classification problem at the sentence or claim level, mainly adopting intrinsic self-checking (Varshney et al., 2023; Luo et al., 2024; Manakul et al., 2023; Cheng et al., 2025; Du et al., 2024; Chen et al., 2024; Bazarova et al., 2025), or tool-augmented verification (Chern et al., 2023; Min et al., 2023; Wei et al., 2024; Dhuliawala et al., 2024; Cheng et al., 2024; Bao et al., 2025; Xie et al., 2025). However, response-level methods suffer from a fundamental granularity limitation: real-world LLM outputs are often only partially hallucinated, and sentence-level judgments cannot localize problematic spans, leading to coarse feedback and unnecessary rejection of correct content.

Span-level hallucination detection aims to localize unsupported or conflicting spans, enabling fine-grained analysis and correction. Due to the lack of high-quality datasets, prior works (Niu et al., 2024; Mishra et al., 2024) primarily focus on dataset construction. Existing datasets can be grouped into manually constructed and automatically constructed. Manually annotated datasets (Niu et al., 2024) provide high-quality labels but are expensive and difficult to scale. Automatically constructed datasets (Mishra et al., 2024) typically synthesize hallucinations via span perturbation: they first generate factually correct claims, and then replace or insert spans into these claims using strong LLMs. While scalable, hallucination claims produced in this way often exhibit surface-level flaws such as disfluency, incoherence, or inconsistency, which differ from the fluent, coherent, and contextually plausible yet factually incorrect hallucinations generated by real-world LLMs, limiting

model generalization to real-world hallucinations. Moreover, most existing datasets (Mishra et al., 2024) only provide span labels or edits without explicit explanations, resulting in limited explainability and reduced practical utility. These limitations highlight the need for an automated dataset construction framework that can generate realistic LLM-style hallucinations while providing explainable span-level annotation, motivating our iterative self-refinement method.

3 Dataset Construction

As shown in Figure 2, our data construction framework consists of three key stages: (1) Hallucination-inducing Edit Suggestion Generation, (2) Hallucinated Claim Generation via Iterative Self-Refinement, and (3) Explainable Hallucination Annotation and Data Filtering. In this section, we first introduce our fine-grained hallucination taxonomy and then elaborate on each stage in detail.

3.1 Fine-Grained Hallucination Taxonomy

In this paper, we focus on factual hallucinations, i.e., generated content that contradicts or unverifiable with the world knowledge (Li et al., 2024; Mishra et al., 2024; Huang et al., 2025), since it has become one of the primary sources of erroneous responses by LLMs. Following Mishra et al. (2024), we operationalize world knowledge as documents from the web that are most relevant to the given statement. We categorize hallucinations into 11 fine-grained types, including Entity, Attribute, Relation, Predicate, Number, Location, Temporal, Omission, Sentence, Subjective, and Unverifiable hallucinations. These categories cover both factual inconsistencies grounded in local evidence (e.g., entity, number, and location errors) and higher-level semantic or context-dependent hallucinations (e.g., omission, sentence-level, and subjective errors). Detailed definitions of each hallucination type are provided in Appendix A.

3.2 Hallucination-inducing Edit Suggestion Generation

The first stage involves three steps: evidence documents collection, factually accurate claim generation, and hallucination-inducing edit suggestion generation. To ensure domain diversity, we first collect documents as reference evidence from 10 knowledge-intensive datasets spanning 8 domains. The detailed information is provided in the Appendix B. We then generate factually accurate claims based on the collected evidence documents. To diversify the types of factually accurate claims, we construct data for two tasks: summarization and question answering. We employ GPT-4o to (i) generate summaries from the given evidence documents, or (ii) perform QA by asking and answering questions grounded in those documents. We treat the generated summaries and answers as factually accurate claims. Finally, based on our hallucination taxonomy and the generated factually accurate claims, we produce structured hallucination-introducing edit suggestions to inject hallucinations into the factually accurate claims. Specifically, we first randomly select five types of hallucination from the pre-defined fine-grained hallucination taxonomy, and then prompt GPT-4o to generate, for each selected type, an edit suggestion that introduces the corresponding type of hallucination into the given factually accurate claim. Each edit suggestion consists of three components: the hallucination type to be introduced, the original span, and the proposed revision to the original span that introduces the corresponding hallucination type.

3.3 Hallucinated Claim Generation via Iterative Self-Refinement

The second stage consists of two components: iterative self-refinement and final hallucinated claim generation.

Iterative Self-refinement. To produce more realistic hallucinations, after Stage 1 generates an initial set of hallucination-introducing edit suggestions, we adopt a self-improvement mechanism that iteratively refines these suggestions. Each iteration consists of two steps: hallucinated text generation and the refinement of the edit suggestions. **Step 1: Hallucinated text generation.** In each iteration, given the current set of edit suggestions, we first apply the current hallucination-introducing edit suggestions to the factually correct text to generate a hallucinated text that includes only the hallucinations specified in the edit suggestions. To ensure that the generated hallucinated claim introduces only the hallucinations mentioned in the suggestions, we prompt GPT-4o with the full set of edit suggestions and the factually accurate claim, asking it to produce a hallucinated claim that only modifies the parts mentioned in the edit suggestions while keeping all other parts identical to the given factually accurate claim. **Step 2: Edit suggestion refinement.** After generating the hallucinated claim, we analyze issues in the gener-

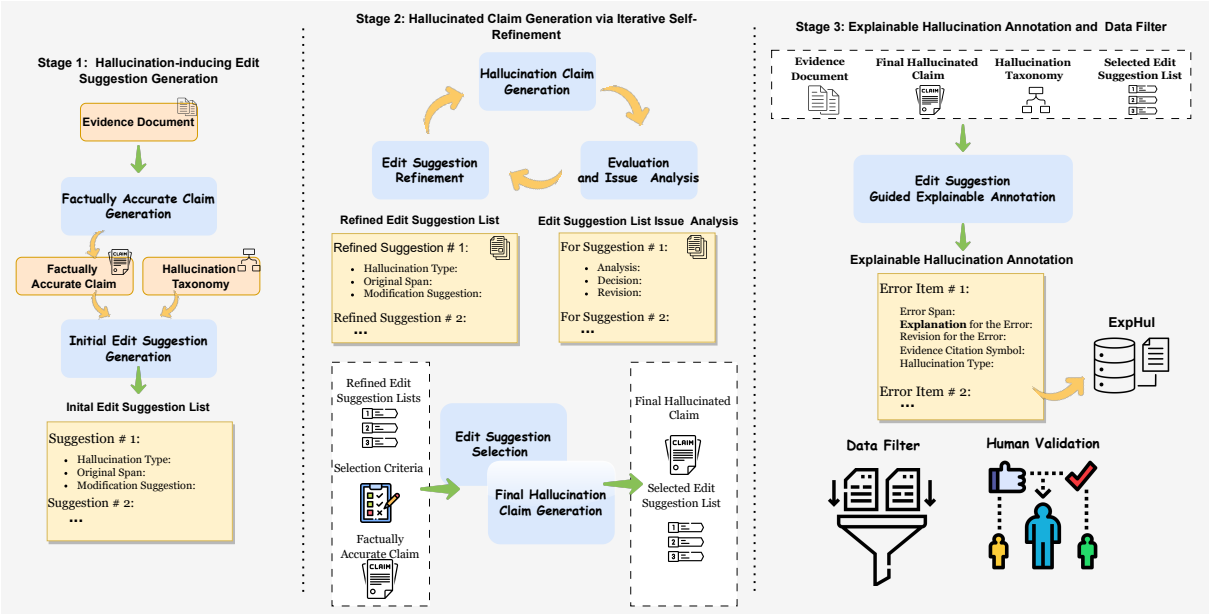

Figure 2: Overview of the ExpHul dataset construction framework. **Stage 1** generates initial structured hallucination-inducing edit suggestions conditioned on the factually correct claim and the fine-grained hallucination taxonomy. **Stage 2** uses a generate–critique–refine loop, refining edit suggestions to produce more realistic hallucinations. **Stage 3** performs explainable annotation guided by the selected edit suggestions, producing hallucination spans with evidence-based explanations, revision suggestions, evidence citations, and types.

ated hallucinated claim and revise or remove any edit suggestions that cause disfluency, incoherence, or inconsistency in the hallucinated claim, resulting in updated hallucination-inducing edit suggestions. Specifically, based on the current hallucinated claim, we prompt GPT-4o first to evaluate each hallucination-introducing edit suggestion to determine whether it leads to disfluency, inconsistency, or incoherence, and provide detailed reasons for each assessment, and then modify or delete problematic edit suggestions according to this analysis, while retaining reasonable ones, yielding an updated list of edit instructions. Finally, after generating the updated list of edit suggestions, we repeat the above refinement process once more to obtain the final candidate list of hallucination-inducing edit suggestions.

Final Hallucinated Claim Generation. After generating the final candidate list of hallucination-inducing edit suggestions, we use GPT-4o to select the most appropriate subset of edit suggestions from the final candidate list and apply them to the factually correct claim from Stage 1 to generate the final hallucinated claim. Specifically, to further select the most suitable eidt suggestions and ensure compatibility among multiple suggestions, we design three evaluation criteria: 1) coherence: The modification should not disrupt the logical flow of the passage or make the passage incoherent; 2) consistency: The modification should fit within the context and not create contradictions; 3) non-conflict: If multiple edit suggestions are selected, they should not conflict with each other. The combined modifications should still result in a coherent passage. According to the criteria, we prompt GPT-4o to select suggestions that do not introduce contextual inconsistency or incoherence in the hallucinated claim and do not conflict with each other, yielding the final set of edit suggestions. Then, we apply the selected edit suggestion(s) to the factually accurate claim produced in Stage 1 to obtain the final hallucinated claim. To ensure that the generated hallucinated claim introduces only the hallucinations specified in the selected suggestions, we prompt GPT-4o with the selected edit suggestions and the factually correct claim, asking it to generate a hallucinated claim that modifies only the parts mentioned in the edit suggestions while keeping all other parts identical to the given factually correct claim.

3.4 Explainable Hallucination Annotation and Data Filtering

The final stage comprises two steps: explainable span-level hallucination annotation and data filtering. **Explainable Span-level Hallucination An-**

notation. In this step, we perform detailed span-level hallucination annotations on the hallucinated claim generated in Stage 2 and provide explanations for each annotated hallucinated span. To further improve annotation accuracy, we use the final hallucination-inducing edit suggestions selected in Stage 2 as additional guidelines to guide GPT-4o in producing hallucination annotations. Specifically, we prompt GPT-4o, given the evidence documents, the annotation guidelines, and the predefined hallucination categories, to identify and annotate all hallucination spans in the given hallucinated claim. For each identified hallucination span, we annotate the following five components: 1) Hallucination Location: the exact text span corresponding to the hallucination; 2) Explanation: an evidence-based justification of why this span is considered hallucinated; 3) Revision suggestion: a proposed revision of the hallucinated span grounded in the evidence; 4) Evidence citation symbol: the index of the sentences in the evidence document used to identify this hallucinated span; 5) Hallucination category: the hallucination category to which this span belongs under our fine-grained taxonomy.

Data Filtering and Test-set Validation. To ensure the quality of the annotated data, we apply a filtering step. We filter instances in which the annotated hallucinations are inconsistent with the hallucinations introduced by the final hallucination-inducing edit suggestions from Stage 2, and retain only those instances where the annotations align, forming the final dataset. To further improve the accuracy of the test-set annotations, we recruited four undergraduate students majoring in computer science to manually verify and correct the test-set data. Each annotator should first check whether the number of annotated hallucination spans in an instance is correct, i.e., whether there are over- or under-annotations, and make necessary corrections. Then, for each annotated hallucination span, the annotator should review and, as needed, revise each component, including: whether the hallucinated span is precise; whether the explanation is reasonable; whether the revision suggestion is appropriate; whether the referenced evidence is correct; and whether the hallucination category is correct.

3.5 Dataset Analysis

Dataset Statistics The ExpHul dataset contains 7,458 samples, including 5,691 hallucinated and 1,767 non-hallucinated instances. In total, 15,555 hallucinated spans are annotated, with an average of 2.73 spans per hallucinated sample, reflecting the frequent co-occurrence of multiple hallucinations within a single hallucinated claim. We split the dataset into 6,944 training samples (14,843 spans) and 514 test samples (712 spans). The test-set are further validated by four undergraduate students. The dataset covers 11 fine-grained hallucination types across 8 domains, enabling comprehensive evaluation of span-level detection, type classification, and explanation and revision generation.

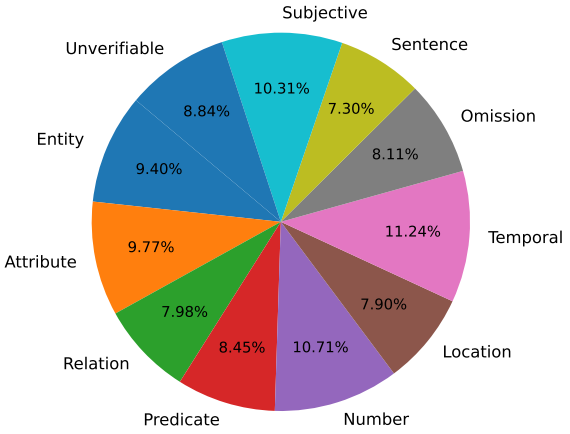

Figure 3: Distribution of Hallucination Types.

Hallucination Type Distribution Figure 3 illustrates the presents the distribution of fine-grained hallucination types in ExpHul. Overall, the dataset exhibits a relatively balanced distribution across the 11 fine-grained hallucination categories, avoiding dominance by any single error type. Temporal (11.24%) and Number (10.71%) hallucinations are the most frequent, while context-dependent errors such as Omission (8.11%) and Unverifiable (8.84%) are also well represented, ensuring diverse and challenging hallucination phenomena.

4 Experiment

4.1 Experimental Setup

Evaluation. We evaluate span-level hallucination detection from three perspectives: **1) Span detection** measures whether hallucinated spans are correctly identified. A predicted span is considered correct if its overlap with a gold span exceeds a predefined threshold, where the overlap is computed using the intersection-over-union (IoU). We report precision, recall, and F1 scores. **2) Token localization** evaluates hallucination span boundary prediction accuracy. We report micro-averaged

token-level precision, recall, and F1 by converting predicted and gold hallucination spans into token sets and computing overlap-based true positives across all samples. **3) Fine-grained type classification** evaluates whether each hallucinated span is correctly classified into a specific fine-grained hallucination category. Precision, recall, and F1 scores are reported.

Compared Methods. We compare our detection model trained on ExpHul with three categories of baselines. **Closed-source LLMs** include GPT-4o (OpenAI et al., 2024), GPT-5.1, GPT-5.2, and Gemini-3.0-Pro. These models are evaluated in a prompt-based manner using a unified instruction and output format. **Open-source LLMs** include LLaMA-3.1-8B-Inst, Qwen2.5-7B/14B/32B-Inst (Qwen et al., 2025), Qwen3-235B-Inst (Yang et al., 2025), and DeepSeek-R1-Distill-Qwen-14B (DeepSeek-AI, 2025). All open-source models are evaluated with the same prompt and output schema as the closed-source models to ensure fair comparison. **Span-level hallucination detection methods** include RAGTruth (Niu et al., 2024) and FAVA (Mishra et al., 2024). RAGTruth is a span-level hallucination detection model trained on a manually constructed dataset; we use the dataset released by the authors and train a LLaMA-3.1-8B model with the same backbone as ours. FAVA is trained on a span-perturbation-based synthetic dataset, and we directly use the model released by the authors.

Implementation Details for the Detection Model. We adopt LLaMA-3.1-8B as the backbone of our detection model and fine-tune it on the training split of ExpHul. The model is trained for one epoch with a batch size of 1, a gradient accumulation step of 8, and an initial learning rate of 2e-5. All experiments are conducted on four NVIDIA A100 GPUs.

4.2 Main Results

Table 1 reports the main results on our dataset, evaluating span-level hallucination detection from three perspectives. Three key observations can be found: First, strong closed-source LLMs struggle with precise hallucination boundary localization and fine-grained type classification. Although they perform competitively at the span level, their performances drop under more fine-grained evaluations, indicating that our benchmark is challenging.

Second, existing span-level hallucination detection methods exhibit limited generalization on our dataset. For instance, FAVA achieves span-level performance to 45.70% but remains far behind both strong LLMs and our model, particularly in token-level localization, highlighting the difficulty of transferring existing methods to our benchmark.

Third, our model achieves competitive performance with strong closed-source models at a lower cost, while consistently outperforming all open-source LLMs and existing baselines. For example, our model achieves a span-level F1 score of 66.76% and a token-level F1 of 53.71%, closely matching Gemini-3.0-Pro (68.29% and 55.07%) and achieves the best fine-grained type F1 score of 40.76%. These results demonstrate that effective span-level hallucination detection can be achieved efficiently when trained on the proposed dataset.

4.3 Detection Difficulty of Different Hallucination Types

To investigate the relative difficulty of detecting different hallucination types, we report category-wise span detection recall for four models: two strong closed-source LLMs (GPT-5.1-medium and Gemini-3-pro), our model trained on the proposed dataset, and FAVA trained on span-perturbation-based hallucination synthetic data. The span detection recall is computed as the proportion of gold hallucination spans of a given category that are detected, regardless of error type correctness, since we focus on analyzing span localization rather than classification accuracy. As shown in Figure 4, three main observations are found:

First, entity-centric hallucinations (e.g., Location, Temporal, and Entity) are relatively easy to detect. Both strong LLMs and our model achieve high recall on these categories, indicating that errors grounded in local concrete factual information can often be identified using primarily local evidence, making them easier to detect.

Second, context-dependent hallucinations, such as Omission and Subjective hallucinations, are harder to detect. Omission hallucinations are the most challenging, with GPT-5.1-medium and Gemini-3-pro below 35% recall, and FAVA only 8.57%. This highlights the inherent difficulty of identifying missing information, which requires reasoning about information completeness rather than detecting explicit inconsistencies. Subjective hallucinations also exhibit large performance gaps and generally low recall for strong LLMs, indicating that detecting unsupported opinions or value judgments demands deeper semantic understanding beyond factual alignment.

Method	Span Detection			Token Localization			Type Classification		
	P	R	F1	P	R	F1	P	R	F1
GPT-4o	36.68	61.52	45.96	18.72	59.14	28.44	29.93	34.73	30.74
GPT-5.1-medium	61.95	68.82	65.20	48.21	56.68	52.10	40.21	39.07	38.02
GPT-5.1-high	59.66	68.54	63.79	46.45	55.87	50.73	40.54	38.70	37.73
GPT-5.2-none	62.59	60.39	61.47	46.87	63.49	53.93	38.34	33.78	34.66
GPT-5.2-medium	56.19	70.08	62.38	41.39	68.90	51.71	39.32	41.88	38.94
Gemini-3.0-Pro	65.97	70.79	68.29	55.15	54.98	55.07	42.47	41.98	40.22
LlaMA-3.1-8B-Inst	6.06	21.49	9.45	5.04	79.72	9.48	2.47	5.88	2.94
Qwen2.5-7B-Inst	9.82	21.49	13.48	6.47	62.93	11.73	14.99	6.39	5.45
Qwen2.5-14B-Inst	24.84	49.44	33.07	12.52	57.96	20.59	15.09	22.43	16.54
DeepSeek-R1-Distill-Qwen-14B	53.68	38.90	45.11	33.03	44.59	37.95	29.94	17.63	19.73
Qwen2.5-32B-Inst	43.03	56.32	48.78	25.80	52.69	34.64	32.45	29.13	26.58
Qwen3-235B-Inst	30.85	58.15	40.31	16.94	71.81	27.41	26.13	32.83	27.46
RAGTruth	54.95	14.04	22.37	30.72	17.16	22.02	–	–	–
FAVA	52.76	40.31	45.70	45.39	31.08	36.90	–	–	–
Ours	64.81	68.82	66.76	42.79	72.14	53.71	39.66	44.99	40.76

Table 1: Span-level hallucination detection performance of closed-source LLMs, open-source LLMs, existing detection methods, and our model, evaluated from three perspectives: hallucinated span detection, span token localization, and fine-grained hallucination type classification.

Third, our model demonstrates robust performance across most categories, remaining competitive with or surpassing LLMs on challenging types such as Subjective and Omission. In contrast, FAVA exhibits consistently low recall across nearly all categories, with particularly poor performance on Omission hallucination. This gap suggests that span-perturbation-based hallucination synthesis is insufficient to capture the diversity and subtlety of real-world hallucinations, further validating the effectiveness of our dataset construction approach.

4.4 Explanation and Revision Evaluation

To evaluate explanation and revision quality, we adopt a reference-based evaluation using GPT-5.1 as an expert evaluator. Given the evidence document, hallucinated claim, ground-truth span, explanation, and revision, GPT-5.1 scores the model-generated explanation and revision for each correctly detected span on a 1–5 scale (1 = completely incorrect or missing; 2 = mostly incorrect with major mismatches or errors; 3 = partially correct with notable issues; 4 = mostly correct with only minor issues; 5 = excellent match). Explanation quality measures whether the explanation correctly identifies why the span is hallucinatory and aligns with the ground truth without factual or logical errors, while revision quality assesses whether the suggested revision appropriately corrects the hallucination, is grounded in the evidence, preserves the reference intent, and introduces no new hallucinations.

Table 2 presents the explanation and revision evaluation results. Three main observations can be drawn: First, strong closed-source LLMs consistently achieve high average explanation and revision scores with low standard deviations, indicating their strong ability in producing reliable explanations and evidence-grounded revisions.

Second, among open-source models, performance improves with increasing model scale. Smaller models such as LlaMA-3.1-8B-Inst struggle to generate reliable explanations and revision suggestions, showing both low average scores and large standard deviations In contrast, larger models such as Qwen2.5-32B-Inst and Qwen3-235B-Inst approach the performance of proprietary models.

Third, despite using the same backbone as LLaMA-3.1-8B-Inst, our model improves both explanation and revision quality, achieving competitive performance with strong closed-source LLMs. For example, our model achieves an average explanation score of 4.85 and an average revision score of 4.80, with lower standard deviations (0.55

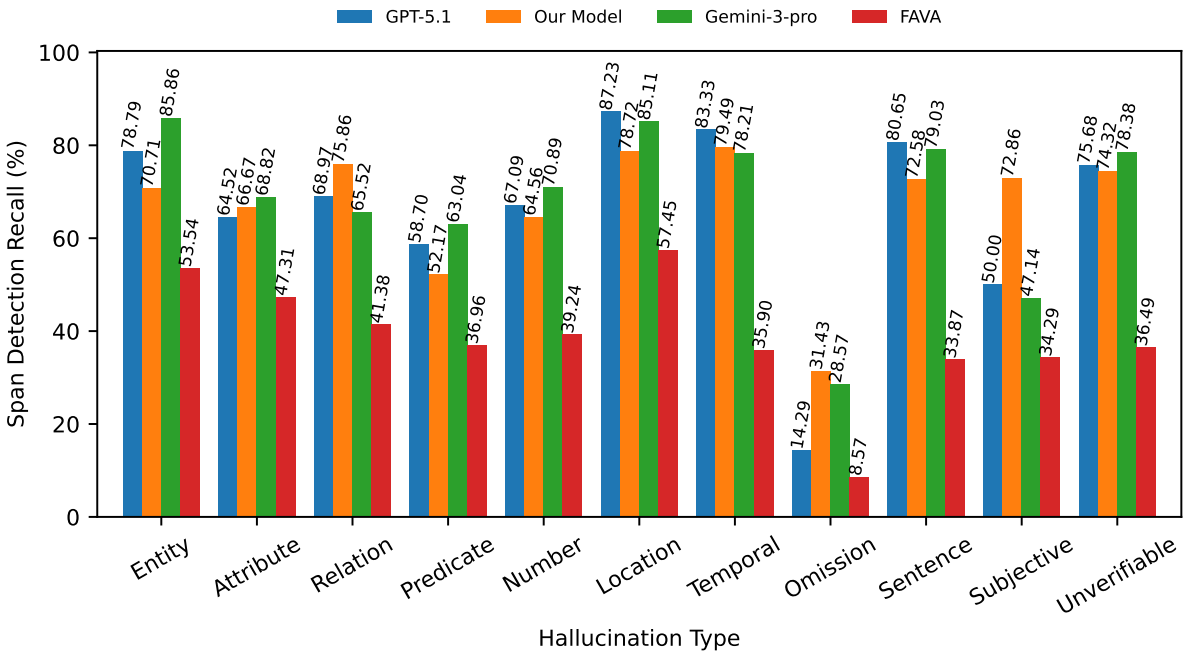

Figure 4: Detection difficulty of different hallucination types. The span detection recall is defined as the proportion of gold hallucination spans of a given category that are detected by the model, independent of error type prediction.

Method	Avg. Explanation	Avg. Revision
GPT-4o	4.92 ± 0.41	4.85 ± 0.48
GPT-5.1-medium	4.94 ± 0.36	4.88 ± 0.45
GPT-5.1-high	4.95 ± 0.34	4.89 ± 0.45
GPT-5.2-none	4.94 ± 0.35	4.90 ± 0.42
GPT-5.2-medium	4.94 ± 0.39	4.88 ± 0.50
Gemini-3.0-Pro	4.96 ± 0.24	4.75 ± 0.67
LlaMA-3.1-8B-Inst	3.74 ± 1.59	3.41 ± 1.63
Qwen2.5-7B-Inst	4.25 ± 1.35	4.21 ± 1.33
Qwen2.5-14B-Inst	4.78 ± 0.66	4.65 ± 0.84
DeepSeek-R1-Distill-Qwen-14B	4.75 ± 0.75	4.55 ± 0.94
Qwen2.5-32B-Inst	4.89 ± 0.54	4.74 ± 0.72
Qwen3-235B-Inst	4.90 ± 0.51	4.76 ± 0.68
Ours	4.85 ± 0.55	4.80 ± 0.62

Table 2: Reference-based evaluation results on explanation quality and revision quality. A large language model is used to score each output on a 1–5 scale, with mean $\pm$ standard deviation reported.

and 0.62), outperforming LLaMA-3.1-8B-Inst by a large margin, indicating that high-quality explanations and corrections can be obtained even with small backbone models when trained on appropriately constructed, explanation-aware data.

Human Evaluation To assess the reliability of LLM-based evaluation, we measure the agreement between LLM-generated scores and human judgments using rank-based correlation metrics. Human annotators assess explanation and revision quality using the same criteria and 1–5 Likert-scale as the LLM evaluator, ensuring a fair comparison. As shown in Table 3, LLM-based evaluation shows strong agreement with human judgments, suggesting that LLM-based scoring provides a reliable proxy for human judgment and is suitable for automatic evaluation in our setting.

Aspect	Spearman's ρ	Kendall's τ_b
Explanation Quality	0.77	0.75
Revision Quality	0.66	0.63

Table 3: Rank correlation between LLM-based evaluation and human judgments over 390 samples (13 models $\times$ 30 samples each).

5 Conclusion

We propose a self-refinement framework that generates realistic hallucinated claims with explainable annotations through structured edit suggestions and a generate–critique–refine loop. Using this framework, we construct the ExpHul dataset and train an 8B-parameter span-level hallucination detector that accurately localizes hallucinated spans and produces high-quality explanations and revisions.

6 Limitations

This work has several limitations that point to directions for future research. First, while leveraging a strong LLM enables the generation of more realistic hallucinated claims and high-quality explainable annotations, it may limit accessibility and reproducibility in resource-constrained settings. Future work could explore using smaller or open-source models to reduce this dependency. Second, although our fine-grained taxonomy covers a broad range of factual hallucinations, it does not exhaust all possible hallucination phenomena. Extending the taxonomy and the construction framework to capture more complex cases remains an open challenge.

References

Zepeng Bao, Shen Zhou, Qiankun Pi, Jianhao Chen, Mayi Xu, Ming Zhong, Yuanyuan Zhu, and Tieyun Qian. 2025. Can a small model learn to look before it leaps? dynamic learning and proactive correction for hallucination detection. *arXiv preprint arXiv:2511.05854*.

Alexandra Bazarova, Aleksandr Yugay, Andrey Shulga, Alina Ermilova, Andrei Volodichev, Konstantin Polev, Julia Belikova, Rauf Parchiev, Dmitry Simakov, Maxim Savchenko, et al. 2025. Hallucination detection in llms with topological divergence on attention graphs. *arXiv preprint arXiv:2504.10063*.

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. 2024. Inside: Llms' internal states retain the power of hallucination detection. In *The Twelfth International Conference on Learning Representations*.

Guanliang Chen, Jie Yang, Claudia Hauff, and Geert-Jan Houben. 2018. Learningq: a large-scale dataset for educational question generation. In *Proceedings of the international AAAI conference on web and social media*, volume 12.

Xiaoxia Cheng, Zeqi Tan, Zhe Zheng, and Weiming Lu. 2025. G2ldetect: A global-to-local approach for hallucination detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 102–109.

Xiaoxue Cheng, Junyi Li, Wayne Xin Zhao, Hongzhi Zhang, Fuzheng Zhang, Di Zhang, Kun Gai, and Ji-Rong Wen. 2024. Small agent can also rock! empowering small language models as hallucination detector. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14600–14615.

I Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, Pengfei Liu, et al. 2023. Factool: Factuality detection in generative ai–a tool augmented framework for multi-task and multi-domain scenarios. *arXiv preprint arXiv:2307.13528*.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A Smith, and Matt Gardner. 2021. A dataset of information-seeking questions and answers anchored in research papers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4599–4610.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Jay DeYoung, Eric Lehman, Benjamin Nye, Iain Marshall, and Byron C Wallace. 2020. Evidence inference 2.0: More data, better models. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, pages 123–132.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2024. Chain-of-verification reduces hallucination in large language models. In *Findings of the association for computational linguistics: ACL 2024*, pages 3563–3578.

Xuefeng Du, Chaowei Xiao, and Sharon Li. 2024. Haloscope: Harnessing unlabeled llm generations for hallucination detection. *Advances in Neural Information Processing Systems*, 37:102948–102972.

Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *NIPS*, pages 1693–1701.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Luyang Huang, Shuyang Cao, Nikolaus Parulian, Heng Ji, and Lu Wang. 2021. Efficient attentions for long document summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1419–1436, Online. Association for Computational Linguistics.

Ziwei Ji, Yuzhe Gu, Wenwei Zhang, Chengqi Lyu, Dahua Lin, and Kai Chen. 2024. Anah: Analytical annotation of hallucinations in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8135–8158.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea

Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577.

Ryo Kamoi, Tanya Goyal, Juan Diego Rodriguez, and Greg Durrett. 2023. Wice: Real-world entailment for claims in wikipedia. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7561–7583.

Yuta Koreeda and Christopher D Manning. 2021. Contractnli: A dataset for document-level natural language inference for contracts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1907–1919.

Junyi Li, Jie Chen, Ruiyang Ren, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2024. The dawn after the dark: An empirical study on factuality hallucination in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10879–10899.

Junyu Luo, Cao Xiao, and Fenglong Ma. 2024. Zero-resource hallucination prevention for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3586–3602.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 9004–9017.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100.

Abhika Mishra, Akari Asai, Vidhisha Balachandran, Yizhong Wang, Graham Neubig, Yulia Tsvetkov, and Hannaneh Hajishirzi. 2024. Fine-grained hallucination detection and editing for language models. In *First Conference on Language Modeling*.

Shashi Narayan, Shay B Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807.

Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, Kashun Shum, Randy Zhong, Juntong Song, and Tong Zhang. 2024. Ragtruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10862–10878.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui

Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunninghman, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Richard Yuanzhe Pang, Alicia Parrish, Nitish Joshi, Nikita Nangia, Jason Phang, Angelica Chen, Vishakh Padmakumar, Johnny Ma, Jana Thompson, He He, et al. 2022. Quality: Question answering with long input texts, yes! In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5336–5358.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Neeraj Varshney, Wenlin Yao, Hongming Zhang, Jianshu Chen, and Dong Yu. 2023. A stitch in time saves nine: Detecting and mitigating hallucinations of llms by validating low-confidence generation. *arXiv preprint arXiv:2307.03987*.

Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Jie Huang, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, et al. 2024. Long-form factuality in large language models. *Advances in Neural Information Processing Systems*, 37:80756–80827.

Zhuohan Xie, Rui Xing, Yuxia Wang, Jiahui Geng, Hasan Iqbal, Dhruv Sahnan, Iryna Gurevych, and Preslav Nakov. 2025. Fire: Fact-checking with iterative retrieval and verification. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 2901–2914.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

A Fine-grained Hallucination Taxonomy

We categorize hallucination into 11 fine-grained types as shown in Table 4.

Hallucination Type	Definition
Entity Error	the entity in a statement contradicts the evidence document.
Attribute Error	a descriptive characteristic (e.g., profession, adjective) of an entity in a statement contradicts the evidence document.
Relation Error	the type of relationship (e.g., familial, social) between two entities contradicts the evidence document.
Predicate Error	the action or state (e.g., verb, condition) described in a statement contradicts the evidence document.
Number Error	the quantity in a statement contradicts the evidence document.
Location Error	the geographical or spatial location in a statement contradicts the evidence document.
Temporal Error	the date, time, or duration in a statement contradicts the evidence document.
Omission Error	crucial parts of the evidence document are omitted from a statement, leading to an incomplete or misleading understanding of the evidence document.
Sentence Error	the overall semantic meaning of the statement contradicts the evidence document, rather than the specific details.
Subjective Error	the expression in a statement is influenced by personal beliefs, opinions, or biases and lacks universal validity, making it inherently subjective.
Unverifiable Error	the full, non-subjective statement cannot be verified or contradicted by the evidence document due to insufficient information or context, making its truthfulness indeterminable.

Table 4: Our fine-grained hallucination taxonomy.

B Evidence Document Collection

To ensure domain diversity, we collect documents as evidence documents from 10 knowledge-intensive datasets spanning 8 domains: biomedicine, Wikipedia, science, news, law, government reports, literature, and education. The source datasets include: PubMedQA (Jin et al., 2019), Evidence Inference (DeYoung et al., 2020), WICE (Kamoi et al., 2023), QASPER (Dasigi et al., 2021), CNN/Daily Mail (See et al., 2017; Hermann et al., 2015), XSum (Narayan et al., 2018), ContractNLI (Koreeda and Manning, 2021), GovReport (Huang et al., 2021), QuALITY (Pang et al., 2022), LearningQ (Chen et al., 2018).

Statistic	Value
Total samples	7,458
Hallucinated samples	5,691
Non-hallucinated samples	1,767
Total hallucinated spans	15,555
Avg. spans per hallucinated sample	2.73
Training samples	6,944
Hallucinated spans (Train)	14,843
Test samples	514
Hallucinated spans (Test)	712
Hallucination types	11
Domains	8

Table 5: Dataset statistics of the ExpHul.

C Dataset Statistics

Table 5 shows the statistics of our dataset.